tulogo.png

Fakultät für Informatik

Professur Rechnernetze und Verteilte Systeme

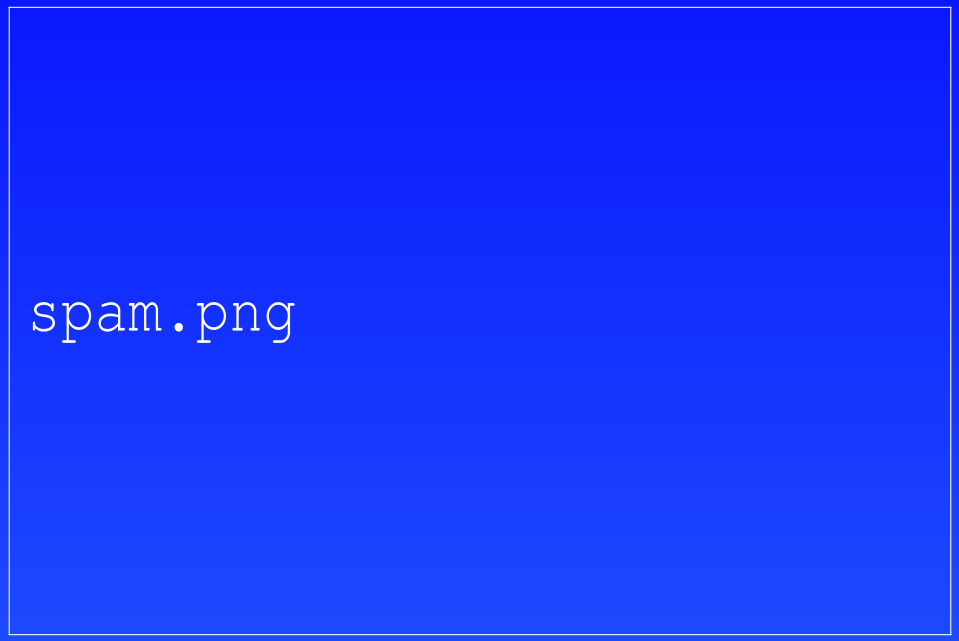
Bayes kontra Spam

Wie ein englischer Mathematiker des 18. Jahrhunderts heute unerwünschte Werbung per E-Mail bekämpft.

Ralph Sontag

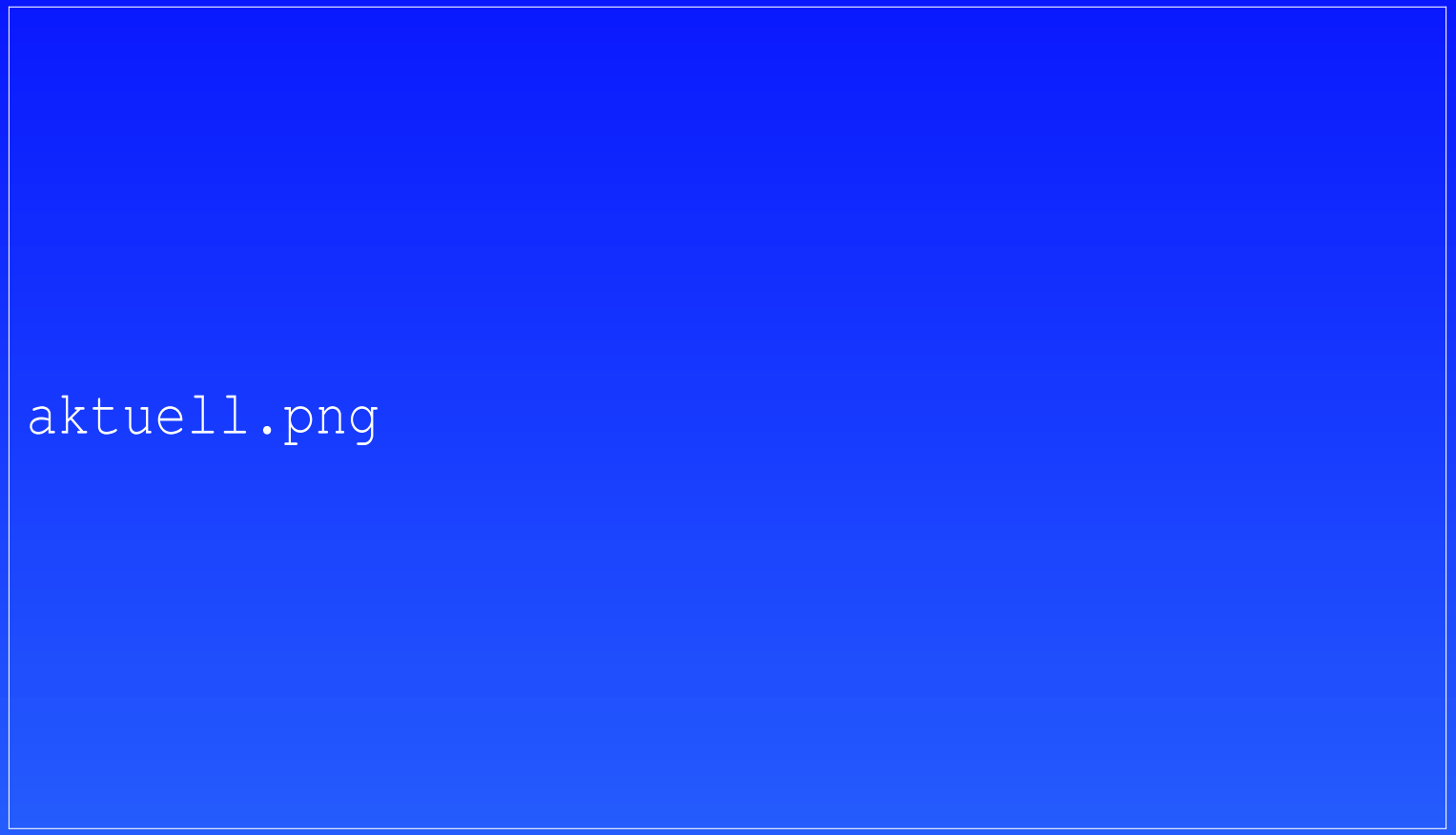
URZ-Workshop
14.-17.4.2003 in Löbsal bei Meißen

Aktuelle Situation



Derzeitige Methoden werden vom Spammer unterlaufen!

Aktuelle Situation



Derzeitige Methoden werden vom Spammer unterlaufen!

Spammersicht

Rücklaufquote:
ca. 15 pro Million (Paul Graham) $\Leftrightarrow 3000/10^6$ bei Mailing

Das ist unglaublich billig!

Das sind ca. 35 Arbeitstage (1 Sekunde pro Mail für's Löschen)!

Ziel wird oft erreicht, sobald die Mail in der Mailbox ist.

Spammersicht

Rücklaufquote:
ca. 15 pro Million (Paul Graham) $\Leftrightarrow 3000/10^6$ bei Mailing

Das ist unglaublich billig!

Das sind ca. 35 Arbeitstage (1 Sekunde pro Mail für's Löschen)!

spam-com.png

Ziel wird oft erreicht, sobald die Mail in der Mailbox ist.

Spammersicht

Rücklaufquote:
ca. 15 pro Million (Paul Graham) $\Leftrightarrow 3000/10^6$ bei Mailing

Das ist unglaublich billig!

Das sind ca. 35 Arbeitstage (1 Sekunde pro Mail für's Löschen)!

spam-com.png

Ziel wird oft erreicht, sobald die Mail in der Mailbox ist.

Spammersicht

Rücklaufquote:
ca. 15 pro Million (Paul Graham) $\Leftrightarrow 3000/10^6$ bei Mailing

Das ist unglaublich billig!

Das sind ca. 35 Arbeitstage (1 Sekunde pro Mail für's Löschen)!

spam-com.png

Ziel wird oft erreicht, sobald die Mail in der Mailbox ist.

Nutzersicht

Zentrale Filter? . . . Wenig Entscheidungsspielraum

Private Filter? . . . Regelverständnis? Regelmäßige Update? Abhängigkeit?

Einzige Chance: Das Ziel des Spammers direkt angreifen!

Spammer wollen etwas „vermarkten“:

- WWW-Seite
- Telefonnummer
- Mailadresse
- Produkt
- . . .

Nutzersicht

Zentrale Filter? . . . Wenig Entscheidungsspielraum

Private Filter? . . . Regelverständnis? Regelmäßige Update? Abhängigkeit?

Einzige Chance: Das Ziel des Spammers direkt angreifen!

Spammer wollen etwas „vermarkten“:

- WWW-Seite
- Telefonnummer
- Mailadresse
- Produkt
- . . .

Nutzersicht

Zentrale Filter? . . . Wenig Entscheidungsspielraum

Private Filter? . . . Regelverständnis? Regelmäßige Update? Abhängigkeit?

Einzige Chance: Das Ziel des Spammers direkt angreifen!

Spammer wollen etwas „vermarkten“:

- WWW-Seite
- Telefonnummer
- Mailadresse
- Produkt
- . . .

Nutzersicht

Zentrale Filter? . . . Wenig Entscheidungsspielraum

Private Filter? . . . Regelverständnis? Regelmäßige Update? Abhängigkeit?

Einzige Chance: Das Ziel des Spammers direkt angreifen!

Spammer wollen etwas „vermarkten“:

- WWW-Seite
- Telefonnummer
- Mailadresse
- Produkt
- . . .

Irgendwann wird nur die „Nutzlast“ ein Kriterium liefern.

Wir erkennen Spam sofort. Kann das ein Programm lernen?

Wir gehen 250 Jahre zurück:

Irgendwann wird nur die „Nutzlast“ ein Kriterium liefern.

Wir erkennen Spam sofort. Kann das ein Programm lernen?

Wir gehen 250 Jahre zurück:

Irgendwann wird nur die „Nutzlast“ ein Kriterium liefern.

Wir erkennen Spam sofort. Kann das ein Programm lernen?

Wir gehen 250 Jahre zurück:

Wie groß ist der Saturn?

Klassische Wahrscheinlichkeitstheorie: Das ist keine Zufallszahl!

Was macht man mit Schätzungen? ⇨ *Bayes*: Intuitiveres Herangehen.

Thomas Bayes: ✻ 1702 (London), ✝ 17. 4. 1761 (Tunbridge Wells)

In an introduction which he has writ to this Essay, he says, that his design at first in thinking on the subject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumstances, upon supposition that we know nothing concerning it but that, under the same circumstances, it has happened a certain number of times, and failed a certain other number of times.

Richard Price über Bayes' „Essay towards solving a problem in the doctrine of chances"

Wie groß ist der Saturn?

Klassische Wahrscheinlichkeitstheorie: Das ist keine Zufallszahl!

Was macht man mit Schätzungen? ⇨ *Bayes*: Intuitiveres Herangehen.

Thomas Bayes: ✻ 1702 (London), ✝ 17. 4. 1761 (Tunbridge Wells)

In an introduction which he has writ to this Essay, he says, that his design at first in thinking on the subject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumstances, upon supposition that we know nothing concerning it but that, under the same circumstances, it has happened a certain number of times, and failed a certain other number of times.

Richard Price über Bayes' „Essay towards solving a problem in the doctrine of chances"

Wie groß ist der Saturn?

Klassische Wahrscheinlichkeitstheorie: Das ist keine Zufallszahl!

Was macht man mit Schätzungen? ⇨ *Bayes*: Intuitiveres Herangehen.

Thomas Bayes: ✻ 1702 (London), ✝ 17. 4. 1761 (Tunbridge Wells)

In an introduction which he has writ to this Essay, he says, that his design at first in thinking on the subject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumstances, upon supposition that we know nothing concerning it but that, under the same circumstances, it has happened a certain number of times, and failed a certain other number of times.

Richard Price über Bayes' „Essay towards solving a problem in the doctrine of chances"

Wie groß ist der Saturn?

Klassische Wahrscheinlichkeitstheorie: Das ist keine Zufallszahl!
Was macht man mit Schätzungen? ⇨ *Bayes*: Intuitiveres Herangehen.
Thomas Bayes: ✻ 1702 (London), ✝ 17. 4. 1761 (Tunbridge Wells)

In an introduction which he has writ to this Essay, he says, that his design at first in thinking on the subject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumstances, upon supposition that we know nothing concerning it but that, under the same circumstances, it has happened a certain number of times, and failed a certain other number of times.
Richard Price über Bayes' „Essay towards solving a problem in the doctrine of chances"

Der Satz von Bayes

$$P(H|D) = \frac{P(D|H) * P(H)}{P(D)} \qquad (I)$$

Der Satz von Bayes

$$P(H|D) = \frac{P(D|H) * P(H)}{P(D)} \qquad (1)$$

$P()$ Wahrscheinlichkeit

Der Satz von Bayes

$$P(H|D) = \frac{P(D|H) * P(H)}{P(D)} \quad (1)$$

$P()$	Wahrscheinlichkeit	
H	Hypothese	Das ist eine gute Mail!

Der Satz von Bayes

$$P(H|D) = \frac{P(D|H) * P(H)}{P(D)} \quad (1)$$

$P()$	Wahrscheinlichkeit	
H	Hypothese	Das ist eine gute Mail!
D	Daten, Ereignis	Die Mail enthält „Money“

Der Satz von Bayes

$$P(H|D) = \frac{P(D|H) * P(H)}{P(D)} \quad (1)$$

$P()$	Wahrscheinlichkeit	
H	Hypothese	Das ist eine gute Mail!
D	Daten, Ereignis	Die Mail enthält „Money“
$P(H\|D)$	Bedingung	Wahrscheinlichkeit, daß die Mail gut ist, wenn sie „Money“ enthält.

Der Satz von Bayes

$$P(H|D) = \frac{P(D|H) * P(H)}{P(D)} \quad (1)$$

$P()$	Wahrscheinlichkeit	
H	Hypothese	Das ist eine gute Mail!
D	Daten, Ereignis	Die Mail enthält „Money“
$P(H\|D)$	Bedingung	Wahrscheinlichkeit, daß die Mail gut ist, wenn sie „Money“ enthält.

Rechte Seite der Formel: Häufigkeiten als sinnvolle Annahmen für die Wahrscheinlichkeit

Der Satz von Bayes

$$P(H|D) = \frac{P(D|H) * P(H)}{P(D)} \qquad (1)$$

$P()$	Wahrscheinlichkeit	
H	Hypothese	Das ist eine gute Mail!
D	Daten, Ereignis	Die Mail enthält „Money“
$P(H\|D)$	Bedingung	Wahrscheinlichkeit, daß die Mail gut ist, wenn sie „Money“ enthält.

Rechte Seite der Formel: Häufigkeiten als sinnvolle Annahmen für die Wahrscheinlichkeit

. . . und praktisch?

Anmerkungen

- Unbekannte Wörter: Hier irrt Graham!

Anmerkungen

- Unbekannte Wörter: Hier irrt Graham!
- Auswerten von mehreren Werten $p_1, \ldots p_n$

Anmerkungen

- Unbekannte Wörter: Hier irrt Graham!
- Auswerten von mehreren Werten $p_1, \ldots p_n$

  Annahme: Unabhängigkeit (ein zweiter mathematischer Fehler :-))

Anmerkungen

- Unbekannte Wörter: Hier irrt Graham!
- Auswerten von mehreren Werten $p_1, \ldots p_n$

  Annahme: Unabhängigkeit (ein zweiter mathematischer Fehler :-))

  Beschränkung auf n (noch ein Fehler ...)

$$P = \frac{\prod_{i=1}^{n} p_i}{\prod_{i=1}^{n} (1 - p_i) + \prod_{i=1}^{n} p_i} \tag{2}$$

- Eindeutige Wörter ⇨ Wahrscheinlichkeit = 1.0 ⇨ Gefahr von Fehlern

Anmerkungen

- Unbekannte Wörter: Hier irrt Graham!
- Auswerten von mehreren Werten $p_1, \ldots p_n$

  Annahme: Unabhängigkeit (ein zweiter mathematischer Fehler :-))

  Beschränkung auf n (noch ein Fehler ...)

$$P = \frac{\prod_{i=1}^{n} p_i}{\prod_{i=1}^{n} (1 - p_i) + \prod_{i=1}^{n} p_i} \tag{2}$$

- Eindeutige Wörter ⇨ Wahrscheinlichkeit = 1.0 ⇨ Gefahr von Fehlern

Besser: Interpretation der Folder als Stichprobe in klassischer Manier ⇨ Maximalwert von 0.05 bzw. 0.95

Besser: Interpretation der Folder als Stichprobe in klassischer Manier ⇨ Maximalwert von 0.05 bzw. 0.95

Noch besser: Abhängig von Anzahl des Vorkommens Maximum festlegen!

Jedenfalls erhalten wir keine Wahrscheinlichkeit, sondern auch nur eine Maßzahl.

Wichtig: Der Nutzer muß konsistent entscheiden!

Wer heute als Spam wegwirft, was er gestern als Ham verdaute, trainiert einen Kalender aber keinen Filter.

Algorithmus: Training

1. E-Mail einordnen (Nutzer!)

Algorithmus: Training

1. E-Mail einordnen (Nutzer!)
2. E-Mail zerlegen

   Wie? (Domains trennen? Wörter trennen?)

Algorithmus: Training

1. E-Mail einordnen (Nutzer!)
2. E-Mail zerlegen

   Wie? (Domains trennen? Wörter trennen?)
3. Tabellen für Spam und Ham aufbauen, Tokens zählen

Algorithmus: Sortieren

1. E-Mail zerlegen

Algorithmus: Sortieren

1. E-Mail zerlegen

2. Wörter in Tabellen suchen

Algorithmus: Sortieren

1. E-Mail zerlegen
2. Wörter in Tabellen suchen
3. Vorschlag: n Extremwerte wählen. Graham: $n = 15$
   (Evtl. Wichtung nötig ⇨ häufige Wörter!)

Algorithmus: Sortieren

1. E-Mail zerlegen
2. Wörter in Tabellen suchen
3. Vorschlag: n Extremwerte wählen. Graham: $n = 15$
   (Evtl. Wichtung nötig ⇨ häufige Wörter!)
4. Wahrscheinlichkeit nach (2) berechnen

Algorithmus: Sortieren

1. E-Mail zerlegen
2. Wörter in Tabellen suchen
3. Vorschlag: n Extremwerte wählen. Graham: $n = 15$
   (Evtl. Wichtung nötig ⇨ häufige Wörter!)
4. Wahrscheinlichkeit nach (2) berechnen
5. Mail einsortieren (Graham: $P > 0.9 \Rightarrow$ Spam)

Algorithmus: Sortieren

1. E-Mail zerlegen
2. Wörter in Tabellen suchen
3. Vorschlag: n Extremwerte wählen. Graham: $n = 15$
   (Evtl. Wichtung nötig ⇨ häufige Wörter!)
4. Wahrscheinlichkeit nach (2) berechnen
5. Mail einsortieren (Graham: $P > 0.9 \Rightarrow$ Spam)
6. Filter justieren, evtl. Nutzerkorrektur beachten

Intermezzo

- Nicht nur Trennung Spam/Ham möglich
- Folder sollten veralten – Änderungen der Spammer auffangen!
- Vermeidung Fehler 1. Art? (Graham: Verdopplung der Wortzahl im Ham-Folder ⇨???)
- White-Lists unnötig, sparen aber Rechenzeit
- Mitlernendes System: „Hallo Ralph“ **war** mal eindeutig Ham
- Risiko: Wortkarge Freunde mit wechselnden Adressen:

> Hallo, bin unterwegs – schau mal auf http:// …– Gruß! R.

Software: Popfile

- Perl-Skript, Plattformunabhängig
- agiert als POP-Proxy:

  Popserver: *localhost*; Nutzer: *org-popserver:nutzer*
- Sortieren in beliebig viele „Buckets“ möglich
- Markierung durch Headerzeile oder im Subject
- Kein „Default“, d.h., die Buckets sind gleichwertig.
- *http://popfile.sourceforge.net/*

Ausblick

Warum erst jetzt?

Ausblick

Warum erst jetzt? Zu viele falsche Positive, zu wenig Erfolg!?

Patrick Pantel and Dekang Lin. „SpamCop – A Spam Classification & Organization Program.“

Ausblick

Warum erst jetzt? Zu viele falsche Positive, zu wenig Erfolg!?

Patrick Pantel and Dekang Lin. „SpamCop – A Spam Classification & Organization Program.“

Ursachen:

- zu wenig Training
- nur Body ausgewertet
- Versuch, Wortstämme zu nutzen
- Auswertung aller Tokens, nicht nur der n signifikantesten ⇨ Fehler bei längerem Spam, leicht auszuhebeln
- keine reine Textanalyse – Email-Strukturen beachten
- Fehler nicht zufällig: Spammer sind aktiv!

Problem: HTML

Ignorieren?

Problem: HTML

Ignorieren? ⇨ Verzicht auf viele wertvolle Hinweise

Parsen?

Problem: HTML

Ignorieren? ⇨ Verzicht auf viele wertvolle Hinweise

Parsen? ⇨ Filter wird zum HTML-Erkenner

Vielleicht nur bestimmte Tags verwerten: *img, bgcolor, font* . . .

Und weiter?

- Analyse von Wortpaaren ⇨ bessere Kontextsensitivität
- Zerlegung von Wörtern (xxxporn ⇨ xxx + porn)
- Unscharfe Erkennung: *M0ney*
- Tokentrennung tunen: Punkte und Kommas zwischen Ziffen nicht als Trenner auffassen: IP-Adressen, Preise . . .
- Headerzeilen extra behandeln: *Subject*Money*

Und weiter?

- Analyse von Wortpaaren ⇨ bessere Kontextsensitivität
- Zerlegung von Wörtern (xxxporn ⇨ xxx + porn)
- Unscharfe Erkennung: *M0ney*
- Tokentrennung tunen: Punkte und Kommas zwischen Ziffen nicht als Trenner auffassen: IP-Adressen, Preise . . .
- Headerzeilen extra behandeln: *Subject*Money*

Subject*FREE	0.9999
free!!	0.9999
To*free	0.9998
Subject*free	0.9782
free!	0.9199
Free	0.9198
Url*free	0.9091
FREE	0.8747
From*free	0.7636
free	0.6546

Und weiter?

- Analyse von Wortpaaren ⇨ bessere Kontextsensitivität
- Zerlegung von Wörtern (xxxporn ⇨ xxx + porn)
- Unscharfe Erkennung: *M0ney*
- Tokentrennung tunen: Punkte und Kommas zwischen Ziffen nicht als Trenner auffassen: IP-Adressen, Preise . . .
- Headerzeilen extra behandeln: *Subject*Money*

Subject*FREE	0.9999
free!!	0.9999
To*free	0.9998
Subject*free	0.9782
free!	0.9199
Free	0.9198
Url*free	0.9091
FREE	0.8747
From*free	0.7636
free	0.6546

Degeneration: *Free!!!!!* ⇨ *Free!!* . . . *Free* ⇨ *free*

Vielen Dank!

Kontakt: Ralph Sontag, *sontag@mathematik.tu-chemnitz.de*

Zum Nachlesen: Paul Graham: „A Plan for Spam“ und „Better Bayesian Filtering“
http://www.paulgraham.com/spam.html